J Comput Aided Mol Des (2017) 31:829–839
DOI 10.1007/s10822-017-0044-3

Structure–reactivity modeling using mixture-based representation of chemical reactions

Pavel Polishchuk[1,2,3] · **Timur Madzhidov**[3] · **Timur Gimadiev**[3,5] · **Andrey Bodrov**[3,4] · **Ramil Nugmanov**[3] · **Alexandre Varnek**[3,5]

Received: 19 September 2016 / Accepted: 23 July 2017 / Published online: 27 July 2017

Abstract We describe a novel approach of reaction representation as a combination of two mixtures: a mixture of reactants and a mixture of products. In turn, each mixture can be encoded using an earlier reported approach involving simplex descriptors (SiRMS). The feature vector representing these two mixtures results from either concatenated product and reactant descriptors or the difference between descriptors of products and reactants. This reaction representation doesn't need an explicit labeling of a reaction center. The rigorous "product-out" cross-validation (CV) strategy has been suggested. Unlike the naïve "reaction-out" CV approach based on a random selection of items, the proposed one provides with more realistic estimation of prediction accuracy for reactions resulting in novel products. The new methodology has been applied to model rate constants of E2 reactions. It has been demonstrated that the use of the fragment control domain applicability approach significantly increases prediction accuracy of the models. The models obtained with new "mixture" approach performed better than those required either explicit (Condensed Graph of Reaction) or implicit (reaction fingerprints) reaction center labeling.

Keywords Chemical reactions · Simplex representation of molecular structure · Condensed graph of reaction · Reaction fingerprints · Rate constant prediction · Mixtures

Electronic supplementary material The online version of this article (doi:10.1007/s10822-017-0044-3) contains supplementary material, which is available to authorized users.

✉ Pavel Polishchuk
pavlo.polishchuk@upol.cz

✉ Timur Madzhidov
timur.madzhidov@kpfu.ru

✉ Alexandre Varnek
varnek@unistra.fr

[1] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czech Republic

[2] A.V. Bogatsky Physico-Chemical Institute of National Academy of Sciences of Ukraine, Odessa, Ukraine

[3] A.M. Butlerov Institute of Chemistry, Kazan Federal University, Kazan, Russia

[4] Department of General and Organic Chemistry, Kazan State Medical University, Kazan, Russia

[5] Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France

Introduction

Structure–property modeling of chemical reactions represents a difficult task because of the complexity issue: any chemical reaction involves several molecular species of two types—reactants and products. The major question concerns the preparation of a descriptor vector encoding a chemical reaction which can serve as an input to a modeling software. Earlier, two different methodologies have been used for this purpose. The first one is based on the explicit consideration of a reaction center identified either manually or automatically using atom-to-atom mapping procedure [1]. This approach has been used in most of reported QSPR studies of reactions. Thus, Gasteiger et al. used some physicochemical parameters (charges, polarizabilities, steric accessibilities, parameters for inductive and resonance effects) for selected atoms and bonds to prepare the models for pK_a for aliphatic carboxylic acids [2] and for kinetics of amide hydrolysis [3]. ISIDA fragment descriptors [4, 5] issued from Condensed Graph of